DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST Q&As

Databricks Certified Professional Data Scientist Exam

Pass Databricks DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-SCIENTIST Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.passapply.com/databricks-certified-professional-data-scientist.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

In unsupervised learning which statements correctly applies?

A. It does not have a target variable

B. Instead of telling the machine Predict Y for our data X, we\\'re asking What can you tell me about X?

C. telling the machine Predict Y for our data X

Correct Answer: AB

Explanation: In unsupervised learning we don\\'t have a target variable as we did in classification and regression.

Instead of telling the machine Predict Y for our data X, we\\'re asking What can you tell me about X?

Things we ask the machine to tell us about

X may be What are the six best groups we can make out of X? or What three features occur together most frequently in X?

QUESTION 2

What is the considerable difference between L1 and L2 regularization?

A. L1 regularization has more accuracy of the resulting model

B. Size of the model can be much smaller in L1 regularization than that produced by L2- regularization

C. L2-regularization can be of vital importance when the application is deployed in resource-tight environments such as cell-phones.

D. All of the above are correct

Correct Answer: B

Explanation: The two most common regularization methods are called L1 and L2 regularization. L1 regularization penalizes the weight vector for its L1-norm (i.e. the sum of the absolute values of the weights), whereas L2 regularization uses its L2-norm. There is usually not a considerable difference between the two methods in terms of the accuracy of the resulting model (Gao et al 2007), but L1 regularization has a significant advantage in practice. Because many of the weights of the features become zero as a result of L1- regularized training, the size of the model can be much smaller than that produced by L2- regularization. Compact models require less space on memory and storage, and enable the application to start up quickly. These merits can be of vital importance when the application is deployed in resource-tight environments such as cell-phones. Regularization works by adding the penalty associated with the coefficient values to the error of the hypothesis. This way, an accurate hypothesis with unlikely coefficients would be penalized whila a somewhat less accurate but more conservative hypothesis with low coefficients would not be penalized as much.

QUESTION 3

You are working on a email spam filtering assignment, while working on this you find there is new word e.g.

HadoopExam comes in email, and in your solutions you never come across this word before, hence probability of this words is coming in either email could be zero. So which of the following algorithm can help you to avoid zero probability?

A. Naive Bayes

B. Laplace Smoothing

C. Logistic Regression

D. All of the above

Correct Answer: B

Explanation: Laplace smoothing is a technique for parameter estimation which accounts for unobserved events. It is more robust and will not fail completely when data that has never been observed in training shows up.

QUESTION 4

You are studying the behavior of a population, and you are provided with multidimensional data at the individual level. You have identified four specific individuals who are valuable to your study, and would like to find all users who are most similar to each individual. Which algorithm is the most appropriate for this study?

A. Association rules

B. Decision trees

C. Linear regression

D. K-means clustering

Correct Answer: D

Explanation: kmeans uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. You can control the details of the minimization using several optional input parameters to kmeans, including ones for the initial values of the cluster centroids, and for the maximum number of iterations. Clustering is primarily an exploratory technique to discover hidden structures of the data: possibly as a prelude to more focused analysis or decision processes. Some specific applications of k-means are image processing^ medical and customer segmentation. Clustering is often used as a lead-in to classification. Once the clusters are identified, labels can be applied to each cluster to classify each group based on its characteristics. Marketing and sales groups use k-means to better identify customers who have similar behaviors and spending patterns.

QUESTION 5

Select the correct statement which applies to logistic regression:

A. Computationally inexpensive, easy to implement knowledge representation easy to interpret

B. May have low accuracy

C. Works with Numeric values

D. Only 1 and 3 are correct

E. All 1, 2 and 3 are correct

Correct Answer: E

Explanation: Depending on the size of the data you are uploading, Amazon S3 offers the following options: Logistic regression Pros: Computationally inexpensive, easy to implement knowledge representation easy to interpret Cons: Prone to underfitting, may have low accuracy Works with: Numeric values^ nominal values

QUESTION 6

Refer to image below

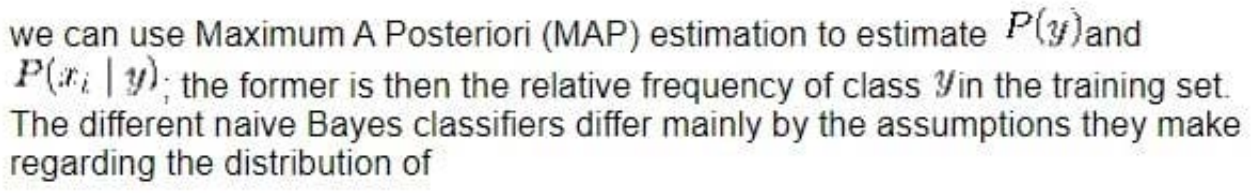
A. Option A

B. Option B

C. Option C

D. Option D

Correct Answer: A



Text

QUESTION 7

You are working on a Data Science project and during the project you have been gibe a responsibility to interview all the stakeholders in the project. In which phase of the project you are?

A. Discovery

B. Data Preparations

C. Creating Models

D. Executing Models

E. Creating visuals from the outcome

F. Operationnalise the models

Correct Answer: A

Explanation: During the discovery phase you will be interviewing all the project stakeholders because they would be having quite a good amount of knowledge for the problem domain you will be working and you also interviewing project sponsors you will get to know what all are the expectations once project get completed. Hence, you will be noting down all the expectations from the project as well as you will be using their expertise in the domain.

QUESTION 8

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn\\'t rain, he incorrectly forecasts rain 10% of the time. Which of the following will you use to calculate the probability whether it will rain on the day of Marie\\'s wedding?

A. Naive Bayes

B. Logistic Regression

C. Random Decision Forests

D. All of the above

Correct Answer: A

Explanation: The sample space is defined by two mutually-exclusive events - it rains or it does not rain. Additionally, a third event occurs when the weatherman predicts rain. You should consider Bayes\\' theorem when the following conditions exist. ?The sample space is partitioned into a set of mutually exclusive events {A1, A2,... :An}. ?Within the sample space, there exists an event B: for which P(B)>; 0. ?The analytical goal is to compute a conditional probability of the form: P (Ak B).

QUESTION 9

Let\\'s say you have two cases as below for the movie ratings

1.

You recommend to a user a movie with four stars and he really doesn\\'t like it and he\\'d rate it two stars

2.

You recommend a movie with three stars but the user loves it (he\\'d rate it five stars). So which statement correctly applies?

A. In both cases, the contribution to the RMSE is the same

B. In both cases, the contribution to the RMSE is the different

C. In both cases, the contribution to the RMSE, could varies

D. None of the above

Correct Answer: A

QUESTION 10

In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model\\'s parameters and the normalizing constant usually ignored in MLEs because:

A. The normalizing constant is always very close to 1

B. The normalizing constant only has a small impact on the maximum likelihood

C. The normalizing constant is often zero and can cause division by zero

D. The normalizing constant doesn\\'t impact the maximizing value

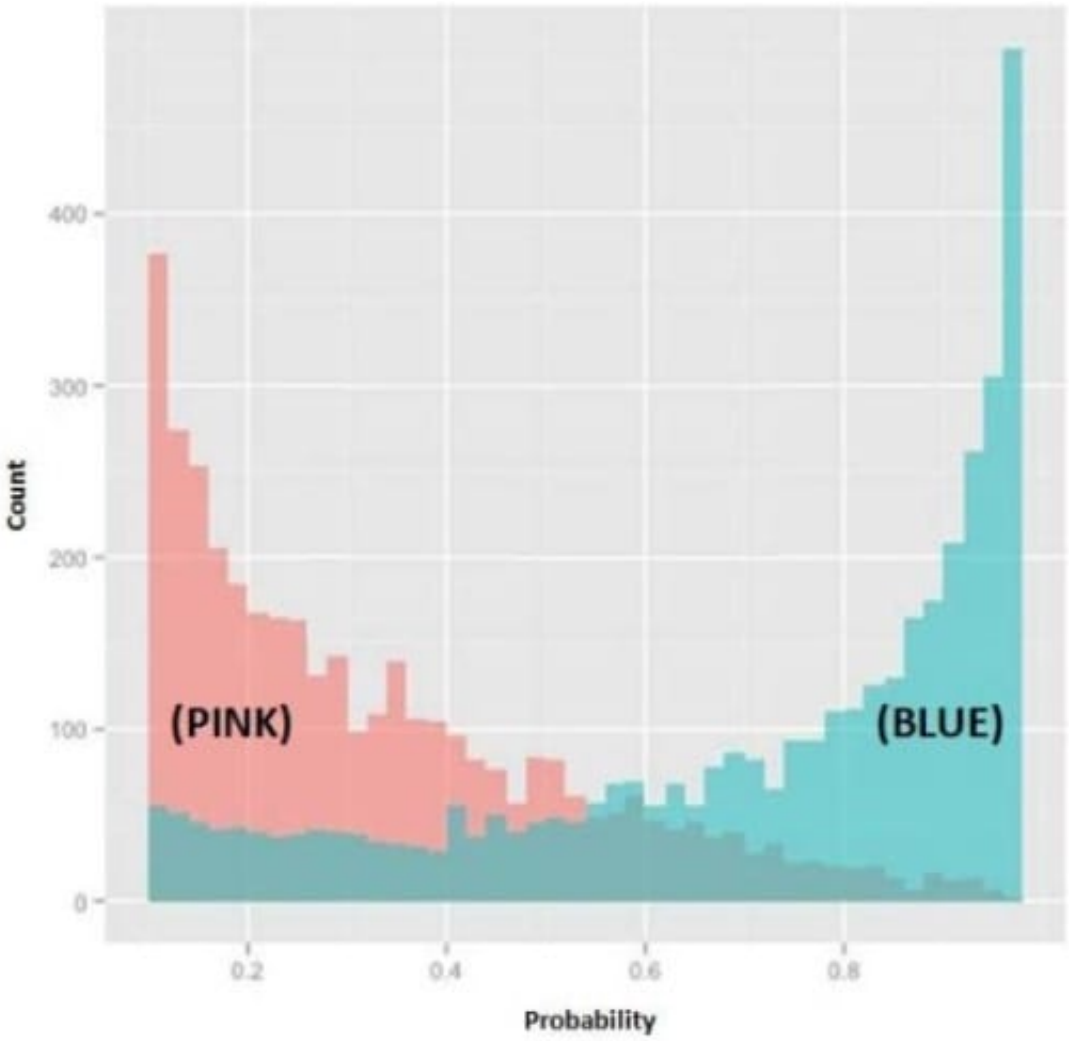
Correct Answer: D

Explanation: (Change the explanation even it is correct)A normalizing constant is positive, and multiplying or dividing a series of values by a positive number does not affect which of them is the largest. Maximum likelihood estimation is

concerned only with finding a maximum value, so normalizing constants can be ignored.

QUESTION 11

Refer to Exhibit



In the exhibit, the x-axis represents the derived probability of a borrower defaulting on a loan. Also in the exhibit, the pink represents borrowers that are known to have not defaulted on their loan, and the blue represents borrowers that are known to have defaulted on their loan. Which analytical method could produce the probabilities needed to build this exhibit?

A. Linear Regression

B. Logistic Regression

C. Discriminant Analysis

D. Association Rules

Correct Answer: B

QUESTION 12

Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (response) variable is binary (0/1); win or lose. The predictor variables of interest are the amount of money spent on the campaign, the amount of time spent campaigning negatively and whether or not the candidate is an incumbent.

Above is an example of:

A. Linear Regression

B. Logistic Regression

C. Recommendation system

D. Maximum likelihood estimation

E. Hierarchical linear models

Correct Answer: B

Explanation: : Logistic regression Pros: Computationally inexpensive, easy to implement, knowledge representation easy to interpret Cons: Prone to underfitting, may have low accuracy Works with: Numeric values, nominal values

QUESTION 13

Which of the following is a Continuous Probability Distributions?

A. Binomial probability distribution

B. Negative binomial distribution

C. Poisson probability distribution

D. Normal probability distribution

Correct Answer: D

QUESTION 14

Refer to the exhibit.

Attribute	Info-Gain
Age	0.0310
Income	0.0100
Gender	0.0034
Credit Score	0.0456

You are building a decision tree. In this exhibit, four variables are listed with their respective values of info-gain. Based

on this information, on which attribute would you expect the next split to be in the decision tree?

A. Credit Score

B. Age

C. Income

D. Gender

Correct Answer: A

QUESTION 15

Assume some output variable "y" is a linear combination of some independent input variables "A" plus some independent noise "e". The way the independent variables are combined is defined by a parameter vector B y=AB+e where X is an m x n matrix. B is a vector of n unknowns, and b is a vector of m values. Assuming that m is not equal to n and the columns of X are linearly independent, which expression correctly solves for B?

A. $b * (A^T * A)^{-1} * A^T$

B. $A^{-1} * b$

C. $(A^T * A)^{-1} * b$

D. $(A^T * A)^{-1} * A^T * b$

A. Option A

B. Option B

C. Option C

D. Option D

Correct Answer: D

Explanation: This is the standard solution of the normal equations for linear regression. Because A is not square, you cannot simply take its inverse.